A Two-Stage Regression Model for Epidemiological Studies With Multivariate Disease Classification Data

Nilanjan CHATTERJEE

Polytomous logistic regression is commonly used to analyze epidemiological data with disease subtype information. In this approach effects of exposures on different disease subtypes are studied through separate exposure odds ratios comparing different case groups to the common control group. This article considers the situation where disease subtypes can be defined using multiple characteristics of a disease. For efficient analysis of such data, a two-stage modeling approach is proposed. At the first stage, a standard polytomous logistic regression model is considered for all possible distinct disease subtypes that can be defined by the cross-classification of the different disease characteristics. At the second stage, the exposure odds ratio parameters for the first-stage disease subtypes are further modeled in terms of the defining characteristics of the subtypes. When the total number of first-stage disease subtypes is small, standard maximum likelihood methods can be used for inference in the proposed model. For dealing with a large number of disease subtypes, a novel semiparametric pseudo-conditional-likelihood approach is proposed that does not require any model assumption about the baseline probabilities for the different disease subtypes. This article develops the asymptotic theory for the estimator and studies its small-sample properties using simulation experiments. The proposed method is applied to study the effect of fiber on the risk of various forms of colorectal adenoma using data available from a large screening study, the Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) Screening Trial.

KEY WORDS: Colorectal adenoma; Genetic marker; Log-linear modeling; Polytomous logistic regression; Protein expression; Pseudo-conditional-likelihood; Semiparametric inference.

1. INTRODUCTION

Diseased subjects in epidemiological studies can often be subtyped using available medical pathological records. Such data, if available, can be used to study "etiologic heterogeneity" among disease subtypes, that is, if the effect of the exposures are different for different disease subtypes. Such findings can be both biologically interesting and have design implications for future studies. For analysis of epidemiologic data with disease subtype information, polytomous logistic regression (Dubin and Pastermack 1986; Hosmer and Lemeshow 1989), which has the same odds ratio parameter interpretation as separate binary logistic regressions comparing subjects of different disease subtypes to the subjects without the disease, is popular among epidemiologists. As better tools for disease classification, including possible use of molecular technologies such as genetic markers (Begg and Zhang 1994; Schroeder and Weinberg 2001) and protein expressions (Terry et al. 2002), are now increasingly available, analytic issues relating to the resulting novel data have become an important area of statistical research. This article addresses the problem of analyzing disease subtype data in epidemiologic studies when the subtypes are defined using multiple characteristics of the disease. To introduce the problem, a motivating example is described first.

A study recently has been completed (Peters et al. 2003) of the association between dietary fiber and prevalent colorectal adenoma, a precursor of cancer, within the large multicenter Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial conducted by the National Cancer Institute. The cases in this study have available pathological data on the various characteristics of the adenomatous polyps. Three specific characteristics focused on were size, villous development, and multiplicity of the adenomatous polyps. The availability of such data posed the problem of whether the adenoma characteristics data can be used to identify certain subtypes of adenoma that may be more strongly related to fiber than other subtypes.

Several issues were raised with this analysis. First was the issue of how to deal with the potentially large number of disease subtypes that can be defined by cross-classification of size, morphology, and multiplicity, especially if one uses the exact size information that was available in the majority of the cases. Use of standard polytomous logistic regression may not be optimal because a study of moderate size may not have enough of all of the different types of diseased subjects to precisely estimate the effect of the covariates on each subtype of disease independent of the other subtypes. A second analytic issue was how to characterize etiologic heterogeneity of the disease subtypes in terms of the defining characteristics of the subtypes. When a number of disease characteristics are being studied simultaneously, it is naturally of interest to determine which characteristics, either individually or jointly, play an important role in defining etiologically heterogeneous disease subtypes.

These problems provided the motivation to propose and study a two-stage regression model as an efficient and systematic way of analyzing epidemiological data with multivariate disease classification information. In this approach, at the first stage, an unstructured polytomous logistic regression approach is used to model the effects of the covariates on all possible disease subtypes that can be defined by cross-classification of the underlying disease characteristics. At the second stage, the subtype-specific regression parameters of the first-stage model are modeled by utilizing the multivariate structure of the subtype definitions and the possibly ordered or continuous nature of certain characteristics. The second-stage model reduces the dimensionality problem associated with the estimation of the first-stage regression parameters. Moreover, the parameters of this model can be interpreted as a measure of etiologic heterogeneity with respect to the defining characteristics of the disease subtypes. It will be shown that the proposed approach is very flexible and can be "semiparametric" in the sense that there is no need to model the baseline disease probabilities for the first-stage disease subtypes, however large the total number of

Nilanjan Chatterjee is Principal Investigator, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD 20852 (E-mail: *chattern@mail.nih.gov*). The author thanks Dr. Richard Hayes for bringing the scientific problem to the author's attention.

Journal of the American Statistical Association
March 2004, Vol. 99, No. 465, Theory and Methods
DOI 10.1198/016214504000000124

these subtypes may be. Inferential techniques have been developed for such semiparametric models.

In Section 2, I propose the two-stage model and focus on the interpretation of the model parameters. In Section 3.1, I briefly describe the standard maximum likelihood (ML) inference procedure in the proposed model. In Section 3.2, I motivate and describe use of an alternative pseudo-conditional-likelihood (PCL) approach that is useful for dealing with a large number of disease subtypes with unspecified baseline disease probabilities. In Section 3.3, I develop the asymptotic theory for the PCL estimator, propose a consistent variance estimator, and show some theoretical connections between the PCL and a semiparametric ML estimator. In Section 4, I study the finite-sample properties of ML and PCL estimators on simulated data. In Section 5, I apply the proposed methods to data from the PLCO study. In Section 6, I discuss some strengths and limitations of the method relative to potential alternative methods and identify some areas of future research.

2. MODEL

Suppose the disease under study can be subtyped using K characteristics. Let the kth characteristic define M_k categories for the disease. One can define a total of $M = M_1 \times M_2 \times \cdots \times M_K$ subtypes based on all possible combinations of the various characteristics. For the ith of N study subjects, let D_i denote the disease status, a polytomous outcome, taking values in $\{0, 1, 2, \ldots, M\}$ with $D_i = 0$, if the ith subject is disease free, and $D_i = m$, if the subject has disease of type m. Let $\mathbf{X}_i$ denote a $P \times 1$ covariate vector associated with the ith subject. Consider an "unstructured" polytomous logistic regression model specified as

$$\Pr(D_i = m|\mathbf{X}_i) = \frac{\exp(\alpha_m + \mathbf{X}_i^T \boldsymbol{\beta}_m)}{1 + \sum_{m=1}^{M} \exp(\alpha_m + \mathbf{X}_i^T \boldsymbol{\beta}_m)}, \qquad m = 1, \ldots, M, \tag{1}$$

where α_m represents the intercept parameter and $\boldsymbol{\beta}_m$ denotes the $P \times 1$ vector of regression coefficients associated with the disease subtype m. In this model, each of the M disease subtypes is treated independently of the others and $\exp(\boldsymbol{\beta}_m)$ can be given the usual covariate odds ratio interpretation for comparing the mth disease group to the nondiseased group.

In the previous approach, even with few covariates and few dimensions for disease characterization, the total number of regression coefficients, given by $Q = M_1 \times M_2 \times \cdots \times M_K \times P$, can easily become very large. The goal here is to propose models with fewer parameters. For the time being, the focus will be on a single covariate. Let $\mathbf{b} = (\beta_1, \beta_2, \ldots, \beta_M)^T$ denote the vector of M regression coefficients corresponding to a single covariate. Because any disease subtype m is defined by a particular combination of K disease characteristics, $\{\beta_m\}_{m=1}^{M}$ can be alternatively indexed as $\{\beta_{i_1 i_2 \ldots i_K}\}_{i_1=1, i_2=1, \ldots, i_K=1}^{M_1, M_2, \ldots, M_K}$. This indexing immediately suggests a linear representation for the log odds ratios as

$$\beta_{i_1 i_2 \ldots i_K} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta^{(1)}_{k_1(i_{k_1})} + \sum_{k_1=1}^{K} \sum_{k_2>k_1}^{K} \theta^{(2)}_{k_1 k_2(i_{k_1} i_{k_2})} + \cdots + \theta^{(K)}_{12\ldots K(i_1 i_2 \ldots i_K)}, \tag{2}$$

where $\theta^{(0)}$ represents the regression coefficient corresponding to a reference disease subtype and the $\theta^{(1)}$'s represent the first-order contrasts, the $\theta^{(2)}$'s represent the second-order contrasts, and so on. A reference level can be chosen for each of the K characteristics and the reference subtype for the disease can be defined jointly by the reference levels of the individual characteristics. For identifiability, all of the $\theta^{(k)}$'s that involve the reference level for any of the K characteristics, excepting $\theta^{(0)}$, need to be set to 0.

Now (2) can be used to specify different models for the regression coefficients by constraining different contrasts to be 0. If all the first-order and higher contrasts are set to 0, the model implies that the effect of the covariate is the same on all of the disease subtypes and $\exp(\theta^{(0)})$ gives the corresponding common covariate odds ratio. If all the second-order and higher contrasts are set to 0, one has the additive model:

$$\beta_{i_1 i_2 \ldots i_K} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta^{(1)}_{k_1(i_{k_1})}. \tag{3}$$

In this model the difference between the covariate effect between two disease subtypes, which have two different levels for the kth characteristic, say i_k and i'_k, but share the same level for all the remaining characteristics, is given by $\beta_{i_1 i_2, \ldots, i_k, \ldots, i_K} - \beta_{i_1 i_2, \ldots, i'_k, \ldots, i_K} = \theta^{(1)}_{k(i_k)} - \theta^{(1)}_{k(i'_k)}$. Thus, model (3) assumes that "etiologic heterogeneity" with respect to one characteristic of a disease does not depend on the other characteristics of the disease. In this case the contrasts of the form $\theta^{(1)}_{k(i_k)} - \theta^{(1)}_{k(i'_k)}$ give a measure of the degree of etiologic heterogeneity (Begg and Zhang 1994) with respect to the kth characteristic.

Model (3) can be further relaxed to consider the following second-order interaction model:

$$\beta_{i_1 i_2 \ldots i_K} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta^{(1)}_{k_1(i_{k_1})} + \sum_{k_1=1}^{K} \sum_{k_2>k_1}^{K} \theta^{(2)}_{k_1 k_2(i_{k_1} i_{k_2})}. \tag{4}$$

In this model the second-degree contrasts of the parameters $\theta^{(2)}_{k_1 k_2}(i_{k_1} i_{k_2})$ measure how the etiologic contrast parameters with respect to the k_1th characteristics are modified by the levels of the k_2th characteristics and vice versa. Inclusion of third-order and higher interaction terms, although possible to include in principle, may not be attractive in practice due to both difficulty of interpretation and lack of parsimony of the model. Higher order models, however, can be useful in testing significance of specific complex interaction terms that are of intrinsic scientific interest.

When the levels of a characteristic induce ordered categories, the etiologic contrast parameters with respect to that characteristic can be further modeled using techniques for modeling association in contingency tables (Anderson 1984; Agresti 1996). Specifically, for ordered characteristics with levels reflecting some kind of progression of the disease, it may be reasonable to assume that the degree of etiologic heterogeneity between any two levels of the characteristic cannot be less than that between two intermediate levels of the same characteristic. If the kth characteristic is ordered and $i_k = 1$ defines the reference level for this characteristic, one convenient way to impose the ordering constraints would be to use the linear regression model:

$$\theta^{(1)}_{k(i_k)} = \theta^{(1)}_k s^{(k)}_{i_k}, \qquad i_k = 1, \ldots, M_k, \tag{5}$$

where $\{0 = s_1^{(k)} \le s_2^{(k)} \le \cdots \le s_{M_k}^{(k)}\}$ is a set of scores assigned to the M_k levels of the characteristic. This model summarizes the degree of etiologic heterogeneity with respect to the kth characteristic into a single regression coefficient $\theta_k^{(1)}$ with $\theta_k^{(1)} = 0$ implying no heterogeneity. The scoring approach can be extended to the second-order interaction model (4) following techniques for modeling interactions in contingency tables (Agresti 1996).

The background for use of scoring merits discussion. Scoring, in general, is a technique for analysis of ordinal categorical data (Agresti 1996). For regression analysis of ordered categorical outcomes, Anderson (1984) proposed the use of scoring to impose ordering constraints on the regression coefficients of unordered polytomous logistic regression models. Moreover, he advocated the use of "one dimensional stereographic modeling" where a single set of scores is used to model regression coefficients for different covariates.

In the context of this article, for continuous characteristics, the exact measurements or some transformation of them can be used as the scores. For truly categorical characteristics, however, the choice of scores may be subjective. For analysis of contingency tables, Graubard and Korn (1987) is a good reference for various choices of scores and their relative merits and demerits. In the context of the polytomous logistic regression model, Greenland (1994) discussed issues related to choice of score for ordinal categorical response. In particular, he argued that equally spaced scores, such as the integer score $s_j = j$, would correspond to an exponential increase for the odds ratios of the form $\exp(\beta_j) = \exp(\beta_1 j) = \exp(\beta_1)^j$ with outcome level spacing j. This can give rise to unrealistically high values for changes in odds ratios between extreme categories of the outcome. In such situations he recommended consideration of some concave transformation of the integer score, such as $s_j = \sqrt{j}$, which would correspond to a less dramatic change for the odds ratio. Based on both practical and theoretical considerations [see condition (A.2) in the Appendix], I came to a similar conclusion for the choice of scores in the proposed two-stage model. Therefore, it is recommended that one avoid functional forms of scores that can give rise to unrealistically high values of odds ratios for extreme levels of an ordered characteristic. For further guidance on strategies for choosing a suitable form of score, the reader is referred to the analysis of the PLCO data (Sec. 5).

So far, this article has concentrated on a single covariate. Some notation is useful in formulating the problem for multiple covariates. Let $\mathbf{b}_p = (\beta_{1p}, \beta_{2p}, \ldots, \beta_{Mp})^T$ denote the vector of regression coefficients for the M disease subtypes associated with the pth covariate. For each of the covariates, one has a model of the form $\mathbf{b}_p = \mathbf{Z}^{(p)}\boldsymbol{\theta}_p$. Here the design matrix $\mathbf{Z}^{(p)}$ relates the coefficients $\mathbf{b}_p$ of the unstructured polytomous regression model (1) to the lower dimensional second-stage regression parameters $\boldsymbol{\theta}_p$. Thus, if one defines $\mathbf{b} = (\mathbf{b}_1^T, \mathbf{b}_2^T, \ldots, \mathbf{b}_P^T)^T$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_P^T)^T$, one has a model of the form $\mathbf{b} = \mathbf{Z}\boldsymbol{\theta}$, where $\mathbf{Z} = \bigoplus_{p=1}^{P} \mathbf{Z}^{(p)}$ is a block-diagonal matrix with $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(P)}$ as the diagonal blocks. In the definition of $\mathbf{b}$, the regression parameters are grouped by covariates. For describing estimation in the previous model, however, reordering $\mathbf{b}$ to order the parameters by disease groups is more convenient. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_M^T)^T$ define such a grouping of the regression parameters. Clearly, the second-stage model can now be represented as $\boldsymbol{\beta} = \mathbf{Z}^*\boldsymbol{\theta}$, where $\mathbf{Z}^*$ is obtained by a reordering of the rows of the design matrix $\mathbf{Z}$. Hereafter, with a slight abuse of notation, $\mathbf{Z}^*$ will be denoted by $\mathbf{Z}$.

3. INFERENCE

3.1 Maximum Likelihood

Maximum likelihood inference in the proposed two-stage model is relatively straightforward when the total number of disease subtypes that is defined at the first stage of the model is not "too large." In the following the notation $\mathbf{I}_a$ will be used to denote the $a \times a$ identity matrix, $\mathbf{1}_b$ to denote the $b \times 1$ unit vector, $\otimes$ to denote the Kronecker product, and $\operatorname{diag}(\mathbf{q})$ to denote the diagonal matrix with $\mathbf{q}$ as the diagonal. Let $\mathbf{X}_M = \mathbf{I}_M \otimes \mathbf{X}$. Let $Y_{im} = I\,(D_i = m)$ denote the indicator of whether the ith subject is of mth disease subtype. Let $\mathbf{Y}_m^T = (Y_{1m}, \ldots, Y_{Nm})$, for $m = 1, \ldots, M$, and $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_M^T)^T$ and define $\mathbf{P} = E(\mathbf{Y}|\mathbf{X})$. Further, let $\mathbf{W} = \mathbf{D} - \mathbf{A}\mathbf{A}^T$ for $\mathbf{D} = \operatorname{diag}(\mathbf{P})$ and $\mathbf{A} = \mathbf{D}(\mathbf{1}_M \otimes \mathbf{I}_N)$. Assume that $(Y_{i1}, \ldots, Y_{iM}, \mathbf{X}_i)$, $i = 1, \ldots, N$, are independently and identically (iid) distributed.

With this notation, the ML score equations and the corresponding expected information matrix for the parameters $\boldsymbol{\theta}$ can be defined as $\mathbf{Z}^T\mathbf{X}_M^T(\mathbf{Y} - \mathbf{P}) = 0$ and $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} = E(\mathbf{Z}^T\mathbf{X}_M^T\mathbf{W}\mathbf{Z}\mathbf{X}_M)$, respectively. An iterative reweighted least squares algorithm can be used to solve the score equations for $\boldsymbol{\theta}$. Define $\mathbf{Y}^{*(t)} = \mathbf{W}^{(t)^{-1}}(\mathbf{Y} - \mathbf{P}^{(t)}) + \mathbf{X}_M\mathbf{Z}\boldsymbol{\theta}^{(t)}$, where $\mathbf{P}^{(t)}$ and $\mathbf{W}^{(t)}$ are obtained from the corresponding formulas for $\mathbf{P}$ and $\mathbf{W}$ evaluated at the current estimate $\boldsymbol{\theta}^{(t)}$. The updated estimate of $\hat{\boldsymbol{\theta}}$ is obtained by the weighted least squares estimate $\hat{\boldsymbol{\delta}}^{(t+1)} = \{\mathbf{Z}^T\mathbf{X}_M^T\mathbf{W}^{(t)}\mathbf{X}_M\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{X}_M^T\mathbf{W}^{(t)}\mathbf{Y}^{*(t)}$.

3.2 Pseudo-Conditional-Likelihood Estimation

The polytomous regression model (1) involves two sets of parameters—the intercept parameters α_m, $m = 1, \ldots, M$, which determine the baseline prevalence of the different disease subtypes, and the odds ratio regression parameters $\boldsymbol{\beta}_m$, $m = 1, \ldots, M$. In this model, α_m, $m = 1, \ldots, M$, can be treated as a set of "nuisance parameters" in the sense that these parameters themselves are not of scientific interest. Thus, a second-stage model for the Pintercept parameters is not of any intrinsic interest. Moreover, as shown in simulation studies in Section 4, underspecification of the intercept parameters using a second-stage model can actually give rise to bias in estimation of the regression parameters of interest. Thus, in this situation, a "semiparametric modeling" approach that restricts the second-stage model specification only to the regression parameters of interest, but leaves the nuisance intercept parameters completely unspecified is very attractive. For a large number of disease subtypes, however, joint maximum likelihood estimation of the lower dimensional second-stage regression parameters of interest and a large number of intercept parameters can become numerically challenging. To overcome this problem, use of a "pseudo-conditional-likelihood" (PCL) method is proposed that depends only on the regression parameters of interest and is free of the intercept parameters.

Let $\mathcal{C}_1$ and $\mathcal{C}_0$ denote the indices for diseased and nondiseased subjects, respectively. For each $i \in \mathcal{C}_1$, consider a "matched set," $\mathcal{S}_i$, consisting of the ith diseased subject and N_0 nondiseased subjects. For each such matched set, $\mathcal{S}_i$, define

$$L_i^c = \Pr\Bigg[D_i = d_i, D_j = 0; j \in \mathcal{S}_i, j \neq i \,\Bigg|\, \bigcup_{k \in \mathcal{S}_i} \{D_k = d_i, D_l = 0; l \in \mathcal{S}_i, l \neq k\} \Bigg], \qquad (6)$$

where d_i denotes the observed disease subtype for the ith diseased subject. In words, (6) can be described as the conditional likelihood of the observed disease configuration of the members of the matched set $\mathcal{S}_i$ given the marginal information that exactly one member of $\mathcal{S}_i$ is diseased with subtype d_i and the remaining members are disease free. Now define the PCL of the data as

$$L_{\mathrm{PCL}} = \prod_{i \in \mathcal{C}_1} L_i^c = \prod_{i \in \mathcal{C}_1} \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta}_{d_i})}{\exp(\mathbf{X}_i^T \boldsymbol{\beta}_{d_i}) + \sum_{j \in \mathcal{C}_0} \exp(\mathbf{X}_j^T \boldsymbol{\beta}_{d_i})}.$$

In the preceding formula the expression for L_i^c, as derived from the model formula (1), is free of the associated intercept parameter α_{d_i}. Although each individual term of L_{PCL} is defined by a conditional probability, L_{PCL} itself is not an exact conditional likelihood in the sense that it cannot be viewed as a probability of the observed data conditional on certain events. An exact conditional likelihood, based on the likelihood of the data conditional on the sufficient statistics for the intercept parameters—the marginal numbers of disease-free and different types of diseased subjects—will also be free of the intercept parameters. Computation of such an exact conditional likelihood, however, in general may be very complex.

The PCL score equations corresponding to the log-linear model parameters $\boldsymbol{\theta}$ are defined by $\partial L_{\mathrm{PCL}}/\partial \boldsymbol{\theta} = 0$. Using the second-stage model formula $\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\theta}$, the score equations can be represented as $\mathbf{Z}^T T_{\boldsymbol{\beta}} = 0$, where $T_{\boldsymbol{\beta}} = (T_{\boldsymbol{\beta}_1}^T, \ldots, T_{\boldsymbol{\beta}_m}^T)^T$ with

$$T_{\boldsymbol{\beta}_m} = \sum_{i \in \mathcal{C}_1} I(D_i = m) \times \left\{ \mathbf{X}_i - \frac{\mathbf{X}_i \exp(\mathbf{X}_i^T \boldsymbol{\beta}_m) + \sum_{j \in \mathcal{C}_0} \mathbf{X}_j \exp(\mathbf{X}_j^T \boldsymbol{\beta}_m)}{\exp(\mathbf{X}_i^T \boldsymbol{\beta}_m) + \sum_{j \in \mathcal{C}_0} \exp(\mathbf{X}_j^T \boldsymbol{\beta}_m)} \right\}.$$

3.3 Asymptotic Properties of PCL

In what follows the asymptotic properties of the PCL estimator will be studied in a unified framework that allows both of the following scenarios: (1) The total number of first-stage disease subtypes is small and/or fixed, and the number of subjects of each disease subtype increases with increasing sample size. This situation arises when each of the disease characteristics under consideration is categorical with a small number of levels. (2) The total number of first-stage disease subtypes is large and/or increases with the sample size, but the number of subjects of each subtype is small/bounded. This situation can arise, for example, when one or more of the disease characteristics are continuous.

Let $M^{(N)}$ be the total number of first-stage disease subtypes for a fixed sample size N. For $m \geq 1$ let $p_{1m}^{(N)} = \Pr^{(N)}(D = m)$ and let $N_{1m}^{(N)}$ be the number of subjects of disease subtype m in the data. Many of the quantities defined later depend implicitly on the sample size N. For notational convenience the superscript (N) will be suppressed from now on. Finally, assume $\dim(\boldsymbol{\theta}_0)$, the number of second-stage regression parameters, is fixed and does not depend on sample size.

For $l = 0, 1, 2$ let $S_m^{(l)} = \sum_{j \in \mathcal{C}_0} \mathbf{X}_j^{\otimes l} \exp(\mathbf{X}_j^T \boldsymbol{\beta}_m)$ and $s_m^{(l)} = \mathrm{E} S_m^{(l)}/N_0$, where $\mathbf{u}^{\otimes l}$, $l = 0, 1, 2$, denotes $\mathbf{1}$, $\mathbf{u}$, and $\mathbf{u}\mathbf{u}^T$, respectively. Define $\mathcal{J}_m = \{s_m^{(2)}/s_m^{(1)} - [s_m^{(1)}/s_m^{(0)}]^{\otimes 2}\}$. Further, consider the partition of the matrix $\mathbf{Z}^T$ as $\mathbf{Z}^T = [\mathbf{Z}_1^T, \ldots, \mathbf{Z}_M^T]$ so that each of the elements $\mathbf{Z}_m^T$, $m = 1, \ldots, M$, is a $\dim(\boldsymbol{\theta}_0) \times P$ matrix. With this notation the asymptotic property of the PCL estimator can be stated as follows.

Proposition 1. Under the regularity conditions (A.1)–(A.4) listed in the Appendix, the following results hold:

(a) The estimating equations $\mathbf{Z}^T T_{\boldsymbol{\beta}} = 0$ have a unique, consistent sequence of solutions, $\{\hat{\boldsymbol{\theta}}_{\mathrm{PCL}}^N\}_{N \geq 1}$.

(b)

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{PCL}}^N - \boldsymbol{\theta}_0) = \mathcal{I}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[I(D_i > 0) \left\{ \mathbf{X}_i - \frac{s_{D_i}^{(1)}}{s_{D_i}^{(0)}} \right\} + I(D_i = 0) \boldsymbol{\Gamma}_i \right] + o_p(1),$$

where

$$\mathcal{I} = \lim_{N \to \infty} \sum_{m=1}^M p_{1m} \mathbf{Z}_m^T \mathcal{J}_m \mathbf{Z}_m \qquad (7)$$

and

$$\boldsymbol{\Gamma}_i = \lim_{N \to \infty} \sum_{m=1}^M p_{1m} \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta}_m)}{s_m^{(0)}} \times \mathbf{Z}_m^T \left\{ \mathbf{X}_i - \frac{s_m^{(1)}}{s_m^{(0)}} \right\}, \qquad (8)$$

assuming the limits exist.

(c) $\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{PCL}}^N - \boldsymbol{\theta}_0) \to N(0, \boldsymbol{\Omega})$ in distribution, where

$$\boldsymbol{\Omega} = \mathcal{I}^{-1} + \mathcal{I}^{-1} \boldsymbol{\Sigma} \mathcal{I}^{-1} \qquad (9)$$

and $\boldsymbol{\Sigma} = \mathrm{Var}\, I(D_1 = 0) \boldsymbol{\Gamma}_1$.

For variance estimation, a plug-in estimator is proposed based on formula (9). Estimate $\mathcal{I}$ as $\mathbf{Z}^T \hat{\mathcal{J}} \mathbf{Z}/N$, where $\hat{\mathcal{J}} = \bigoplus_{m=1}^M N_{1m} \hat{\mathcal{J}}_m$ with $\hat{\mathcal{J}}_m$ being the plug-in estimator for $\mathcal{J}_m$. Estimate $\boldsymbol{\Sigma}$ by the empirical variance estimator $\hat{\boldsymbol{\Sigma}} = 1/N \times \sum_{j \in \mathcal{C}_0} \hat{\boldsymbol{\Gamma}}_j \hat{\boldsymbol{\Gamma}}_j^T$, where $\hat{\boldsymbol{\Gamma}}_j$ is obtained from formula (8) with p_{1m} replaced by N_{1m}/N, $s_m^{(l)}$ replaced by $\hat{S}_m^{(l)}/N$, and $\boldsymbol{\beta}_m$ replaced by $\hat{\boldsymbol{\beta}}_m$.

Note on Efficiency. In simulation studies (Table 1) when the efficiency of the PCL was compared to that of the ML, in most situations, the PCL was found to have similar efficiency as that of an ML estimator that allows saturated intercept parameters. This empirical observation, as pointed out by a reviewer, motivated a study of the theoretical connection between the ML with saturated intercept model and the PCL method. This gave rise to some interesting alternative insight into the PCL method.

For a fixed value of $\boldsymbol{\beta}$, the saturated ML estimate of α_m, denoted by $\hat{\alpha}_m(\boldsymbol{\beta})$, will satisfy the score equation

$$\exp(\alpha_m) = N_m \left[\sum_{i=1}^N \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta}_m)}{1 + \sum_{m=1}^M \exp(\alpha_m + \mathbf{X}_i^T \boldsymbol{\beta}_m)} \right]^{-1}. \qquad (10)$$

Table 1. Simulation Results for a "Small" Number of Disease Subtypes Defined by Three Dichotomous Characteristics

		Random sample			Case–control sample		
Estimator	*Parameter*	*Mean*	*SE*	*est(SE)*	*Mean*	*SE*	*est(SE)*
ML (correct)	$\theta^{(0)}$	.354	.131	.132	.354	.153	.152
	$\theta^{(1)}_{1(2)}$	.152	.149	.151	.154	.152	.153
	$\theta^{(1)}_{2(2)}$	−.001	.153	.151	−.001	.155	.153
	$\theta^{(1)}_{3(2)}$	.502	.155	.152	.509	.161	.157
ML (underspecified)	$\theta^{(0)}$	.278	.149	.139	.278	.167	.157
	$\theta^{(1)}_{1(2)}$	.218	.139	.143	.221	.142	.145
	$\theta^{(1)}_{2(2)}$	.085	.141	.142	.087	.144	.144
	$\theta^{(1)}_{3(2)}$	.516	.148	.148	.524	.154	.153
ML (unspecified)	$\theta^{(0)}$	.354	.131	.132	.354	.153	.152
	$\theta^{(1)}_{1(2)}$	.152	.149	.151	.154	.152	.153
	$\theta^{(1)}_{2(2)}$	−.001	.153	.151	.000	.155	.153
	$\theta^{(1)}_{3(2)}$	.502	.155	.152	.509	.161	.157
MPCL	$\theta^{(0)}$	.353	.131	.132	.352	.158	.155
	$\theta^{(1)}_{1(2)}$	.152	.149	.152	.157	.157	.157
	$\theta^{(1)}_{2(2)}$	−.001	.152	.152	−.001	.159	.156
	$\theta^{(1)}_{3(2)}$	.502	.157	.153	.518	.180	.169

NOTE: The parameters $\theta^{(0)}$, $\theta^{(1)}_{1(2)}$, $\theta^{(1)}_{2(2)}$, and $\theta^{(1)}_{3(2)}$ correspond to the second-stage model for the regression coefficients of X. The true values of these parameters are .35, .15, 0, and .5, respectively.

Let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^T$. The ML score equation for $\boldsymbol{\theta}$ is of the form $\mathbf{Z}^T\mathbf{R}_{\boldsymbol{\beta}} = 0$, where $\mathbf{R}_{\boldsymbol{\beta}} = (\mathbf{R}^T_{\boldsymbol{\beta}_1}, \dots, \mathbf{R}^T_{\boldsymbol{\beta}_M})^T$ with

$$\mathbf{R}_{\boldsymbol{\beta}_m} \equiv \mathbf{R}_{\boldsymbol{\beta}_m}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{X}_i^T \{I(D_i = m) - P_m(\mathbf{X}_i)\}, \qquad (11)$$

and $P_m(\mathbf{X}_i) = \Pr(D_i = m|\mathbf{X}_i)$ for $m = 1, \dots, M$ is defined in (1). Let $P_0(\mathbf{X}_i) = \Pr(D_i = 0|\mathbf{X}_i) = [1 + \sum_{m=1}^{M} \exp(\alpha_m + \mathbf{X}_i^T\boldsymbol{\beta}_m)]^{-1}$ and note that $P_m(\mathbf{X}_i) = \exp(\alpha_m)\exp(\mathbf{X}_i^T\boldsymbol{\beta}_m)P_0(\mathbf{X}_i)$. Substitution of (10) into (11) now yields the function

$$\mathbf{R}_{\boldsymbol{\beta}_m}\{\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta}\}$$
$$= \sum_{i=1}^{N} I(D_i = m)$$
$$\times \left\{ \mathbf{X}_i - \frac{\sum_{j=1}^{N} \mathbf{X}_j \exp(\mathbf{X}_j^T\boldsymbol{\beta}_m) P_0^{\{\hat{\alpha}(\boldsymbol{\beta}),\boldsymbol{\beta}\}}(\mathbf{X}_j)}{\sum_{j=1}^{N} \exp(\mathbf{X}_j^T\boldsymbol{\beta}_m) P_0^{\{\hat{\alpha}(\boldsymbol{\beta}),\boldsymbol{\beta})\}}(\mathbf{X}_j)} \right\}, \qquad (12)$$

where $P_0^{\{\hat{\alpha}(\boldsymbol{\beta}),\boldsymbol{\beta}\}}(\mathbf{X})$ denote $P_0(X)$ evaluated at $\{\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta}\}$. Now consider a variation of (12), ignoring the dependency of P_0 on $\{\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta}\}$ and instead fixing P_0 at its true value $P_0^{(0)} \equiv P_0^{(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}$. Of course, because $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ is unknown $P_0^{(0)}$ cannot be evaluated, but the sums of the form $\sum_{j=1}^{N} \mathbf{X}_j^{\otimes l} \exp(\mathbf{X}_j^T\boldsymbol{\beta}_m)P_0^{(0)}(\mathbf{X}_j)$ can be empirically estimated by $\sum_{j=1}^{N} I(D_j = 0)\mathbf{X}_j^{\otimes l}\exp(\mathbf{X}_j^T\boldsymbol{\beta}_m)$. It is now easy to see that this approximation to $\mathbf{R}_{\boldsymbol{\beta}_m}\{\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta}\}$ precisely gives $T_{\boldsymbol{\beta}_m}$, the PCL score function for $\boldsymbol{\beta}_m$, except for some asymptotically ignorable terms. The preceding calculations show some inherent connection between PCL and ML estimators and, hence, give some insight into the observed high efficiency of the PCL estimator in simulation studies.

4. SIMULATION EXPERIMENTS

In this section, the finite-sample properties of ML and PCL estimators are studied on simulated data. Data were first simulated in a scenario where the total number of disease subtypes is "small." Three disease characteristics were considered, each with two levels. This defines a total of $2^3 = 8$ disease subtypes. In each simulation, a single exposure variable X was generated following a standard normal distribution. Given the exposure value, the polytomous logistic regression model (1) was used to generate a multinomial outcome with nine cells, one for the nondiseased subjects and eight for subjects of different disease subtypes. It was assumed that the eight intercept parameters followed the second-order interaction model (4) and the eight regression parameters followed the additive model (3). In this additive model the regression parameters $\theta^{(0)}$, $\theta^{(1)}_{1(2)}$, $\theta^{(1)}_{2(2)}$, and $\theta^{(1)}_{3(2)}$ were chosen to be .35, .15, 0, and .5, respectively. In the interaction model for the intercept terms, $\theta^{(0)}$, $\theta^{(1)}_{1(2)}$, $\theta^{(1)}_{2(2)}$, $\theta^{(1)}_{3(2)}$, $\theta^{(2)}_{12(22)}$, $\theta^{(2)}_{13(22)}$, and $\theta^{(2)}_{23(22)}$ were chosen to be −3.84, −.7, −.7, −.7, .5, .5, and .5, respectively. The marginal probability of the disease corresponding to these parameters is approximately 10%. In each replication 2,000 random samples were generated from the preceding model. Also considered was a case–control sample from the same model that selects all the cases (200 on average) obtained from the random sample and a random sample of the controls of the same size as the number of cases. To investigate the effects of underfitting and overfitting the intercept parameters on the ML estimation of the regression parameters, during analysis of each dataset, in addition to fitting the correct model for the intercepts, two other models were also fitted: the additive model (3) and a saturated model that allows eight separate intercept parameters.

Table 1 shows the mean and standard errors of the ML and maximum PCL (MPCL) estimates of the second-stage regression parameters as well as the mean of the corresponding standard error estimates obtained from 1,000 simulated datasets. One can make the following key observations:

1. The bias of the ML estimates with the correctly specified and the saturated intercept model and that of the MPCL estimates are negligibly small. Underspecification of the intercept model, however, produces significant bias in the ML estimates of the regression parameters.
2. The standard errors of the correct and the saturated ML were identical. This result is somewhat counterintuitivebecause one would expect the correctly specified ML to be more efficient than the overfitted ML. Further investigation suggested that this phenomenon is related to the fact that the true second-order interaction model involves seven parameters, only one parameter less than the saturated model.

   Additional simulations were also considered where the correct model for the intercept parameter was an additive model involving only four parameters. Even in this scenario, where some differences between the standard errors of ML correct and ML saturated started to become evident, the relative efficiency of ML saturated always remained very high (at least 90%) compared to that of ML correct. From these simulations it is obvious that one generally does not gain much efficiency for estimation of the regression parameters by specifying a lower order model for the intercept parameters.
3. MPCL has very high efficiency, in most cases 100%, relative to the saturated ML. The maximum loss of efficiency was observed for the case–control sampling design for the parameter $\theta^{(1)}_{3(2)}$. The true value of this parameter was .5, which reflected a very strong effect.
4. The standard inverse information matrix variance estimator for ML and the proposed standard error estimator for MPCL perform very well in estimating the true standard errors of the corresponding estimates.

In the next experiment data were generated from a model where the number of possible disease subtypes is large, but the number of subjects of each particular subtype is small. Three characteristics for the disease, each with six ordered levels, denoted by $i_k = 1, \ldots, 6$, for $k = 1, 2, 3$, were considered. This defines a total of $6^3 = 216$ first-stage disease subtypes. As before, a single exposure variable X was considered following a standard normal distribution. It was assumed that the exposure odds ratio parameters for the 216 disease subtypes followed the additive model (3), where the parameters $\theta^{(1)}_{k(i_k)}$, $i_k = 1, \ldots, 6$, for each of the $k = 3$ characteristics are further specified using a set of scores using (5). For simplicity, it was assumed that the set of scores for the three characteristics are the same, defined by $s_{k(i_k)} = (i_k - 1)^{.3}$ and $s_{k(1)} = 0$. The baseline common regression parameter ($\theta^{(0)}$) for model (3) was chosen to be .35 and the slope parameters in model (5) for the three characteristics, denoted by $\theta^{(1)}_1$, $\theta^{(1)}_2$, and $\theta^{(1)}_3$, were chosen to be .15, 0, and .5. The intercept parameters were allowed to be saturated.

Two hundred and sixteen free intercept parameters were generated from the model

$$\alpha_{i_1 i_2 i_3} = \theta^{(0)} + \sum_{k=1}^{3} \theta^{(1)}_{k(i_k)} + \epsilon_{i_1 i_2 i_3}, \qquad (13)$$

where $\theta^{(0)}$ was chosen to be -2.95, $\theta^{(1)}_{k(i_k)}$, $i_k = 1, \ldots, 6$, were chosen to be $-2.0, -1.6, -1.0, -1.6, -2.0, -2.7$, respectively, for each of the $k = 3$ characteristics, and $\{\epsilon_{i_1 i_2 i_3}\} \overset{\text{iid}}{\sim} N(0, .5^2)$. The marginal probability of the disease corresponding to this set of parameter values was approximately 10%. The set of intercept parameters was generated once and then treated as fixed as data were repeatedly simulated from the corresponding 217-category polytomous regression model. As in Table 1, in each replication, data were generated from the model using both a random sample ($N = 2{,}000$) and a case–control sample (approximately 200 cases and 200 controls). In each replication the PCL method was applied to estimate the parameters for the second-stage model of the regression coefficients, namely, the baseline parameter $\theta^{(0)}$ and the slope parameters $\theta^{(1)}_1$, $\theta^{(1)}_2$, and $\theta^{(1)}_3$. The scores for the categories of the different characteristics were assumed to be known. Computation of the saturated ML with 216 intercept parameters was unstable and was not pursued.

Table 2. Simulation Results for a "Large" Number of Disease Subtypes Defined by Three Six-Level Ordered Characteristics

	Random sample			*Case–control sample*		
Parameter	*Mean*	*SE*	*est(SE)*	*Mean*	*SE*	*est(SE)*
$\theta^{(0)}$	.354	.376	.370	.355	.407	.392
$\theta^{(1)}_1$	.153	.178	.178	.161	.189	.186
$\theta^{(1)}_2$	−.005	.178	.171	−.006	.190	.180
$\theta^{(1)}_3$	.504	.198	.195	.520	.219	.210

NOTE: The parameters $\theta^{(0)}$, $\theta^{(1)}_1$, $\theta^{(1)}_2$, and $\theta^{(1)}_3$ correspond to the second-stage model for the regression coefficients of X. The true values of these parameters are .35, .15, 0, and .5, respectively.

Table 2 shows that both the MPCL estimates and the proposed standard error estimates unbiasedly estimate the respective population parameters. This result is consistent with the asymptotic results that PCL is a valid estimator even in situations where the number of first-stage disease subtypes is very large, but the number of subjects of each subtype is sparse.

5. DATA EXAMPLE

In this section the proposed method is applied to examine the association between dietary fiber and the prevalence of colorectal adenoma using data from the PLCO trial. Besides the main exposure of interest, Fiber (grams), the study also included total calorie intake (Energy), a covariate that is commonly adjusted for examining nutrient–disease association; smoking history (Smoking—Yes/No), a known risk factor for colorectal adenoma; and Age and Gender. Data on 1,755 cases and 18,945 controls were available, where the cases and the controls were defined as subjects with and without adenoma, respectively, as detected during baseline screening examination. Data on adenoma characteristics consisted of the number of adenomatous polyps (single vs. multiple), the presence of any polyp with villous element (Morphology—Yes/No), and the size of the largest polyp (<10 mm or not). In addition, the exact sizes (in millimeters) of the largest polyps were available for 1,628 cases.

First, consider an analysis based on the dichotomous size information. The subscripts 1, 2, and 3 will be used to denote the characteristic size, morphology, and multiplicity, respectively. The first level for each of these characteristics, that is, small for size, nonvillous for morphology, and single for multiplicity, will be labeled as "1," and the second level will be labeled as "2." Eight disease subtypes are defined based on size, morphology, and multiplicity. The eight intercept parameters are left unspecified and different models are considered for the covariate odds ratios. Because the number of disease subtypes was relatively small, the full maximum likelihood procedure was implemented for parameter estimation.

A model for each covariate odds ratio was selected using the following forward selection procedure. Begin with the additive model (3) and test for significance of interaction terms using likelihood ratio tests. For each pair of characteristics, test for the significance of the corresponding interaction term in each of the five covariate odds ratio models. For all significance tests $\alpha = .05$ was chosen as the critical value.

Following this procedure, the interaction between morphology and multiplicity ($\theta^{(2)}_{23(22)}$) in the odds ratio model for Gender was found to be significant. As a final test of goodness of fit, the selected model was compared against a saturated model that allows eight separate odds ratio parameters for each of the covariates and the corresponding likelihood ratio test was found to be insignificant (p value $= .3$). Adenomas with a single, small, and nonvillous polyp, the most common form of adenoma observed in the data, were chosen as the baseline disease subtype.

The following conclusions can be made from the results shown in Table 3: (1) The estimates of $\theta^{(0)}$ reveal that the prevalence of the reference subtype of adenoma was positively associated with lower intake of fiber, higher intake of energy, past smoking history, age, and being male. (2) Estimates of $\theta^{(1)}_{1(2)}$ reveal that fiber–adenoma association is significantly (p value $= .022$) stronger for large adenomas than for small adenomas. Quantitatively, the odds ratio associated with 10 g of fiber for large adenomas was 18% (95% CI: 3–30%)—computed as $100 \times \{1 - \exp(10 \times \hat{\theta}^{(1)}_{1(2)})\}$—smaller than that for the small adenomas. (3) Estimates of $\theta^{(1)}_{2(2)}$ reveal that all the covariates have a similar effect on villous and nonvillous adenomas. (4) Estimates of $\theta^{(1)}_{3(2)}$ suggest that past history of smoking was more strongly associated with risk of multiple adenoma than with risk of single adenoma. (5) The significance of $\theta^{(2)}_{23(22)}$ (p value $= .006$) implies that the male–female difference in risk is stronger for adenomas with villous and multiple polyps than for the other forms of the disease.

Figure 1 compares the ML estimates and 95% confidence intervals for the parameters of the first-stage model—the subtype-specific odds ratios for fiber—from the fitted additive model with those from a saturated polytomous logistic regression model that allows a completely independent effect of the covariates on the eight different adenoma subtypes. It was found that, for most disease subtypes, the saturated and the modeled ML estimates were in close agreement. The confidence intervals for the latter estimates, however, were generally smaller. The biggest difference between the point estimates was observed for adenomas characterized by a single, villous, small polyp (denoted as 010 in the figure). For this type of adenoma, the saturated model estimated the effect of fiber to be harmful, that is, the log odds ratio was slightly less than 0, although the effect was statistically insignificant. In contrast, the two-stage model, which assumes a certain degree of similarity between the subtype-specific exposure odds ratios, "shrinks" this outlying estimate toward the other seven estimates and brings it into the positive range.

Next, consider the use of the exact (continuous) size information that was available on the subset of 1,628 cases. Figure 2 shows the distribution of the adenoma cases by their exact size. Although imperceptible to the eye due to the scaling of the histograms, for each combination of morphology and multiplicity, there were at least 15 adenomas of size 20 mm or larger. Altogether, adenomas of 27 different sizes ranging from 1 mm to 50 mm were observed.

The PCL methodology was applied to an analysis of these data as the total number of distinct adenoma subtypes in this example was large ($M = 27 \times 2 \times 2 = 108$). Similar to Table 3, an additive second-stage model was selected for Fiber, Energy, Smoking, and Age and a single parameter interaction model for Gender. $\theta^{(1)}_{1(s)}$, the contrast between the log exposure odds ratios for adenomas with size s and that for adenomas with a reference size s_0, was modeled using the linear model of the form $\theta^{(1)}_{1(s)} = \theta^{(1)}_1 \times f(s)$ for some fixed score function f. Because the natural integer score function $f(s) = s$ could give rise to unrealistically high values of exposure odds ratios for large adenomas [see the discussion on choice of scores in Sec. 2 and the regularity condition (A.2) in the Appendix], the transformation $f(s) = \log(s)$ was used as a candidate for the score function. As a diagnostic for adequateness of the log transformation, the class of Box–Cox transformations was considered:

$$f_\alpha(s) = \begin{cases} \dfrac{s^\alpha - 1}{\alpha} & \text{if } \alpha \neq 0 \\ \log(s) & \text{if } \alpha = 0. \end{cases}$$

For a given value of α, a measure of goodness of fit for the corresponding score function $f_\alpha(s)$ was defined as the value of the $-2\log(L_{\mathrm{PCL}})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ that maximizes the L_{PCL} for the given score function $f_\alpha(s)$. Based on a crude grid search over α, it was found that the "best fitting transformation" corresponds to $\alpha = -.1$. Thus, considering easy interpretation as well as goodness of fit, the log transformation ($\alpha = 0$) seems

Table 3. PLCO Data—Estimates (standard errors) of the Second-Stage Parameters Using Dichotomized Size Information

Covariate	$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{2(2)}$	$\theta^{(1)}_{3(2)}$	$\theta^{(1)}_{23(22)}$
Fiber	−.022 (.005)	−.019 (.009)	.010 (.010)	−.004 (.008)	
Energy	.018 (.006)	.004 (.010)	−.010 (.012)	.004 (.010)	
Smoking	.240 (.072)	.094 (.118)	−.024 (.135)	.269 (.119)	
Age	.019 (.007)	.005 (.011)	.010 (.013)	.002 (.011)	
Female	−.381 (.081)	−.057 (.132)	.216 (.167)	−.098 (.148)	−.845 (.301)

NOTE: The subscripts 1, 2, and 3 correspond to size, villous status, and multiplicity, respectively.

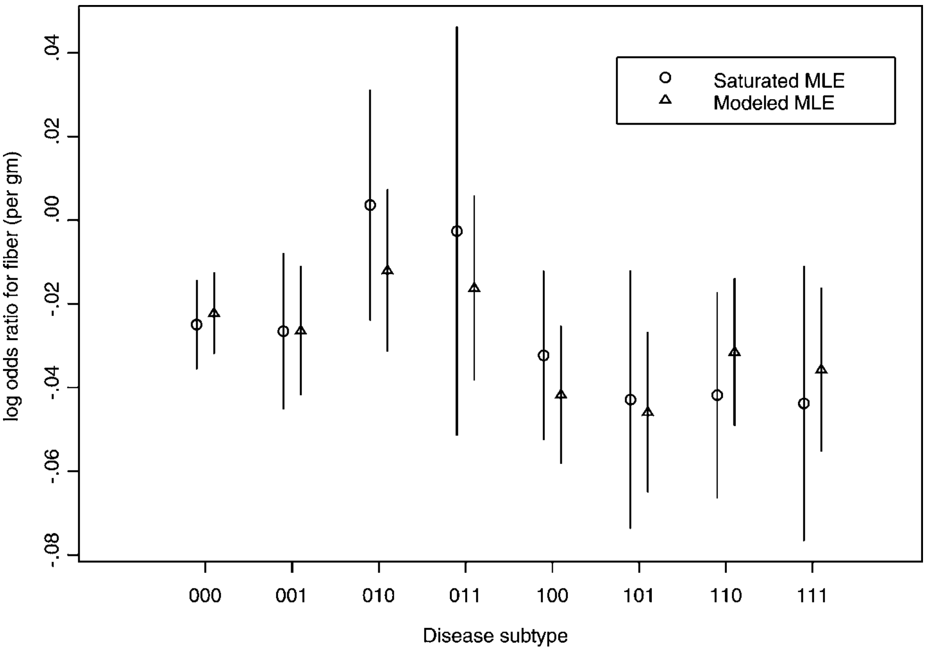

Figure 1. PLCO Data—Estimates of the First-Stage Regression Parameters Using Dichotomized Size Information. The estimates of the subtype-specific log odds ratios were obtained using a fitted two-stage model (triangles) and a saturated polytomous logistic regression model (circles) that allows independent effect of the covariates on each adenoma subtype. The bars around the estimates show the respective 95% confidence intervals. Adenoma subtypes are coded as 000 = (small, nonvillous, single), 001 = (small, villous, multiple), . . . , 111 = (large, villous, multiple).

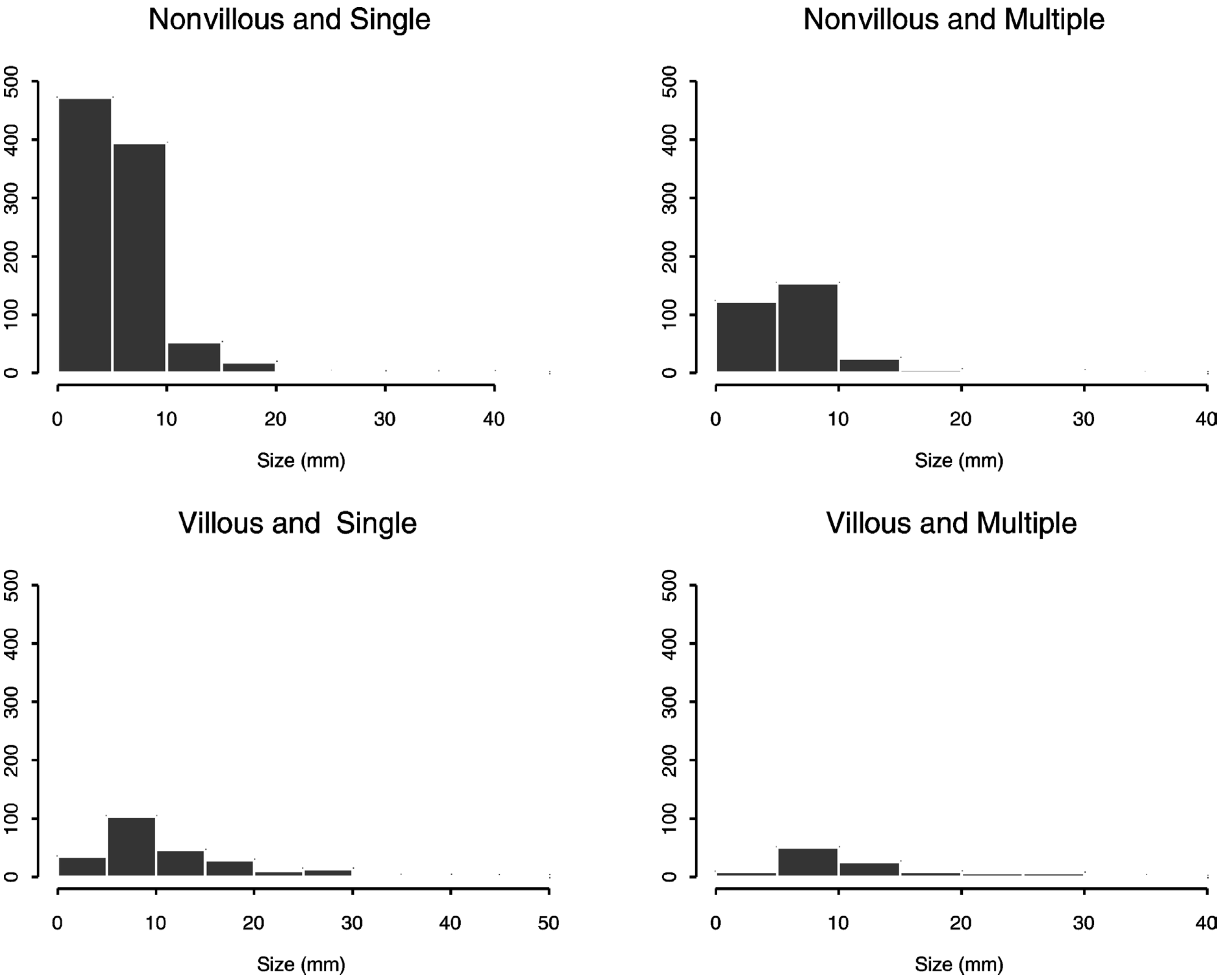

Figure 2. Distribution of Adenoma Cases by Size of Polyps.

Table 4. PLCO Data—Estimates (standard errors) of the Second-Stage Parameters Using Continuous Size Information

Covariate	$\theta^{(0)}$	$\theta_1^{(1)} \times 10$	$\theta_{2(2)}^{(1)}$	$\theta_{3(2)}^{(1)}$	$\theta_{23(22)}^{(1)}$
Fiber	−.025 (.005)	−.013 (.006)	.006 (.010)	−.001 (.009)	—
Energy	.018 (.006)	−.002 (.008)	−.003 (.012)	.001 (.010)	—
Smoking	.278 (.069)	.068 (.092)	−.035 (.138)	.259 (.123)	—
Age	.024 (.007)	−.015 (.009)	.021 (.013)	.008 (.011)	—
Female	−.366 (.079)	−.190 (.103)	.304 (.172)	−.146 (.154)	−.866 (.319)

NOTE: The subscripts 1, 2, and 3 correspond to size, villous status, and multiplicity, respectively.

to be a reasonable choice for "scoring" adenomas of different sizes. Finally, to make the analysis comparable to Table 3, where small adenoma (<10 mm) was defined as the reference level for size, the reference value of size in the continuous analysis was chosen to be 6 mm, the median size for small adenomas in these data. Thus, the score function $\log(s)$ was centered as $\log(s) - \log(6)$ so that $\theta_{1(6)}^{(1)} = 0$.

The estimates shown in Table 4 reveal that continuous analysis of size gives very similar results as those for the categorical analysis (Table 3). In particular, the estimate of θ_0 reveals that Fiber was associated with a significant decreased prevalence of the reference subtype of adenoma. Moreover, the estimate of θ_1 suggests that among adenoma cases higher fiber consumption was negatively associated with the size of the adenoma.

Although various analyses of the data show a consistent association between fiber intake and adenoma size, the results need to be cautiously interpreted due to the cross-sectional nature of the data. Because the latency of an adenoma would be a strong determinant of its size, the association between fiber and size observed here could merely be a reflection of the association between fiber and the latency of the adenomas. Analyses of incident adenoma cases are needed in the future to examine if fiber truly has a larger protective effect on larger adenomas than on smaller adenomas.

6. DISCUSSION

The simulation study (Table 1) showed that underspecification of the nuisance intercept parameters can cause substantial bias in estimation of the regression parameters of interest. Moreover, a lower order model for the intercept parameters, even when correct, does not yield much efficiency gain for estimation of the regression parameters. Given these empirical observations and the fact that the intercept parameters themselves are not of scientific interest, it seems the best strategy would be to leave the intercept parameters completely unspecified. In contrast, the regression parameters are of direct scientific interest and modeling of the first-stage parameters using the lower dimensional second-stage parameters seems attractive for both efficiency and interpretation purposes. Underspecification of the second-stage model for the regression parameters can also create substantial bias in estimates of the first-stage regression parameters. To minimize the possibility of such bias, it is important to use proper model selection methods to choose a parsimonious and yet adequate second-stage model. Within the framework described here, standard model selection techniques, such as forward or backward selection methods, can be used to select the best fitting second-stage model.

Two alternative methods for fitting the proposed model to the data were considered in this article: standard maximum likelihood (ML) and the novel pseudo-conditional-likelihood(PCL). The main advantage of PCL is that it can be computationally simpler than ML when dealing with a large number of disease subtypes. Moreover, based on appropriate asymptotic theory, it was shown that PCL is a valid estimator for the second-stage regression parameters of interest in a semiparametric setting where the baseline disease probabilities are allowed to be unspecified, however large the number of first-stage disease subtypes may be.

At this point it is hard to give a general guideline about how large the number of disease subtypes has to be before the computation of ML becomes difficult and PCL becomes advantageous. However, note that the PCL methodology is valid for both a small and a large number of disease subtypes. Moreover, based on both simulation studies and some theoretical arguments, it was shown that PCL can be thought of as a computationally simple but efficient approximation to an ML estimator that allows saturated intercept parameters. Thus, PCL generally seems to be a very attractive method for analyzing the data irrespective of whether there are a large or small number of disease subtypes. Use of ML, however, can be advantageous if there are a large number of cases with missing disease characteristic information. In this case an EM algorithm–based ML method can be used to efficiently incorporate cases with missing disease characteristic information into the analysis. Similar techniques for handling missing data are not yet available for PCL and are currently under investigation.

A unique feature of the data in the problem presented here is the mixed nature of the outcome variable, defined by a single stratum for the nondiseased subjects and by a multivariate outcome for the diseased subjects. An alternative to the approach given in this article for analyzing such mixed multivariate data could be as follows. First analyze the case–control data using a standard logistic regression model to obtain estimates for the effects of the covariates on overall risk of the disease, irrespective of the subtypes. Data on disease characteristics can then be further analyzed by regression methods for multivariate outcome data (see, for example, sec. 6.3 of McCullagh and Nelder 1989), and the presence of any association between the covariates and the disease characteristics can be taken as an indication of etiologic heterogeneity between the disease subtypes with respect to the corresponding characteristics. At this stage one can use either marginal models that only require specification of the first- and possibly second-order moments of the multivariate data (Zhao and Prentice 1990; Liang, Zeger, and Quaqish

1992) or one can use multivariate models that require specification of a full joint distribution of the multivariate data. Although marginal regression models can be useful for examining the marginal association between the individual disease characteristics and the covariates, the approach may not be suitable if multiple characteristics of a disease jointly define etiologically distinct subtypes of the disease. Moreover, even if the individual disease characteristics completely determine the etiologic heterogeneity among the disease subtypes, the marginal approach may not give a way to combine the estimates of the etiologic contrasts or association parameters corresponding to the individual characteristics to yield estimates of the subtype-specific covariate odds ratios of interest (see Fig. 1).

Specifying a full multivariate model for the disease characteristics, on the other hand, may be a very complex task, depending on the nature of the characteristics, and may require strong distributional assumptions. In contrast, note that the two-stage polytomous regression approach proposed here can handle continuous and categorical characteristics, both ordered and unordered, in a unified fashion and the PCL methodology described here for estimation in this model allows semiparametric inference with minimal distributional assumptions about the underlying characteristics.

Another important advantage of the method proposed here comes into play if the cases are differentially sampled into the study based on their disease subtypes, a situation that may arise either by design or by chance. Due to the multiplicative intercept structure of the first-stage regression model, the proposed method can handle complex case selection designs without changing the parameter interpretations and inference techniques for the regression parameters of interest.

An important feature of the method proposed in this article is that the degree of etiologic heterogeneity with respect to one characteristic is defined by holding the levels of the other characteristics constant. Whether this conditional interpretation of the etiologic contrast parameters is desirable can be debated in certain situations. In the adenoma data, for example, because the variable Size is defined as the size of the largest polyp, it seems natural to condition on multiplicity to examine if the covariates have different effects on adenomas of different sizes. Conditioning on size, on the other hand, to test if the covariates have different effects on adenomas with different multiplicities seems somewhat artificial. In such situations the etiologic contrast parameters in the models should be cautiously interpreted. With this caution in mind, given its various advantages, the proposed methodology overall seems a promising approach to the problem studied in this article.

APPENDIX: REGULARITY CONDITIONS AND PROOF FOR PROPOSITION 1

A.1 Regularity Conditions for Proposition 1

Following are a set of regularity conditions that are sufficient for the conclusion of Proposition 1 to be valid. In these conditions, the quantities p_{1m}, $\boldsymbol{\beta}_m$, $\mathbf{Z}_m$, and $\mathcal{J}_m$ all implicitly depend on the sample size N. For notational convenience, however, the superscript (N) is suppressed.

(A.1) The total probability of being a case, $\sum_{m=1}^{M} p_{1m}$, is fixed at $p_1 \in (0, 1)$.

(A.2) The elements of the design matrix $\mathbf{Z}$ remain uniformly bounded in absolute value by a constant.

(A.3) $\lim_{N\to\infty} \sum_{m=1}^{M} p_{1m}\mathbf{Z}_m^T \mathcal{J}_m \mathbf{Z}_m$ exists and is positive definite.

(A.4) $0 < \mathrm{E}(e^{\mathbf{u}^T\mathbf{X}}) < \infty$ for all $\mathbf{u} \in \Re^P$.

Some discussion of condition (A.2) is warranted. This condition will be trivially satisfied for the models described in this article for unordered characteristics because in this case the matrix $\mathbf{Z}$ simply represents a design matrix of 0s and 1s. For ordered characteristics with fixed scores, $\mathbf{Z}$ will include the scores for different characteristics as its elements. Thus, condition (A.2) in this case is equivalent to requirement of bounded scores for ordered characteristics.

A.2 Asymptotic Theory

First, note that, for purposes of studying asymptotic theory, the score equations $\mathbf{Z}^T T_{\boldsymbol{\beta}} = 0$ can be slightly modified to $\mathbf{Z}^T T_{\boldsymbol{\beta}}^* = 0$, where $T_{\boldsymbol{\beta}}^* = (T_{\boldsymbol{\beta}_1}^{*T}, \ldots, T_{\boldsymbol{\beta}_M}^{*T})^T$ with $T_{\boldsymbol{\beta}_m}^* = \sum_{i\in\mathcal{C}_1} I(D_i = m)\{\mathbf{X}_i - S_m^{(1)}/S_m^{(0)}\}$. The following lemma states a key step that is needed for the proof of Proposition 1.

Lemma 1.

(a)

$$\sqrt{N}\mathbf{W}_{Nm} \equiv \sqrt{N}\left\{\frac{S_m^{(1)}}{S_m^{(0)}} - \frac{s_m^{(1)}}{s_m^{(0)}}\right\}$$

$$\xrightarrow{d} \frac{1}{\sqrt{N}}\frac{1}{s_m^{(0)}}\sum_{i=1}^{N} I(D_i = 0)\exp(\mathbf{X}_i^T\boldsymbol{\beta}_m)\left\{\mathbf{X}_i - \frac{s_m^{(1)}}{s_m^{(0)}}\right\}.$$

(b) The preceding convergence result holds uniformly for all m.

Proof of Lemma 1. Let $\mathcal{Q}$ be the set of all probability measures for $\mathbf{X}$ defined on $\Re^P$ and define the functional $\Psi_m : \mathcal{Q} \mapsto \Re^P$ as $\Psi_m(Q) = \Psi_m^{(1)}(Q)/\Psi_m^{(0)}(Q)$, where $\Psi_m^{(l)}(Q) = \int X^{\otimes l}\exp(\mathbf{X}^T\boldsymbol{\beta}_m)\,dQ(\mathbf{X})$, $l = 0, 1$. With this notation, $\sqrt{N}\mathbf{W}_{Nm}$ can be expressed as $\sqrt{N}[\Psi_m\{Q_N\} - \Psi_m\{Q_0\}]$, where Q_0 denotes the true underlying probability distribution for $\mathbf{X}$ given $D = 0$ and Q_N denotes the corresponding empirical distribution function. The representation of $\sqrt{N}\mathbf{W}_{Nm}$ as a functional of the empirical process Q_N suggests the use of modern empirical process theory to study its asymptotic property. First, by invoking the functional delta theorem (thm. 20.8 of van Der Vaart 1996), one can write $\sqrt{N}[\Psi_m\{Q_N\} - \Psi_m\{Q_0\}] = \sqrt{N}\dot{\Psi}_m[Q_N - Q_0] + o_p(1)$, where $\dot{\Psi}_m : \mathcal{Q} \mapsto \Re^P$ is a continuous linear map that represents the Hadamard derivative (see sec. 20.2 of van Der Vaart 1996 for a definition) of functional $\Psi_m : \mathcal{Q} \mapsto \Re^P$. Because Ψ_m is defined as a ratio of two linear maps, the existence of the Hadamard derivative $\dot{\Psi}_m$ follows by the chain rule of Hadamard differentiability (thm. 20.9 of van Der Vaart 1996). Further, it follows that $\sqrt{N}\dot{\Psi}_m[Q_N - Q_0]$ can be computed as an ordinary derivative given by

$$\sqrt{N}\frac{\delta}{\delta\epsilon}\Psi_m\{(1-\epsilon)Q_0^{(0)} + \epsilon Q_N\}\bigg|_{\epsilon=0}$$

$$= \frac{\sqrt{N}}{\Psi_m^{(0)}(Q_0)}\mathrm{E}_{Q_N}\exp(\mathbf{X}^T\boldsymbol{\beta}_m)\left\{\mathbf{X} - \frac{\Psi_m^{(1)}(Q_0)}{\Psi_m^{(0)}(Q_0)}\right\},$$

which is precisely the expression on the right side of (a) and, hence, part (a) of the lemma is proved.

To prove part (b) of Lemma 1, first note that, by (A.2), there exists a compact set $\mathcal{B} \in \Re^P$ so that $\boldsymbol{\beta}_m \in \mathcal{B}$ for all m in an open neighborhood of $\boldsymbol{\theta}_0$. Thus, in an open neighborhood of the true parameter value $\boldsymbol{\theta}_0$, the function $\exp(\boldsymbol{\beta}_m^T\mathbf{X})$ can be uniformly bounded by a function of the form $\exp(\mathbf{X}^T\boldsymbol{\gamma})$ for some constant vector $\boldsymbol{\gamma}$, uniformly for all m. Now, by ordinary application of the central limit theorem, it can be easily shown that the $\sqrt{N}[\Psi_m^{(l)}(Q_N) - \Psi_m^{(l)}(Q_0)]$, $l = 0, 1$, converge in distribution to appropriate normal distributions with the remainder term going to 0 in probability uniformly in m. Using this, together with the fact that $E_Q \exp(\mathbf{X}^T \boldsymbol{\beta}_m)$ is bounded below uniformly for all m and all Q in an open neighborhood of Q_0 [by (A.2) and (A.4)], one can now easily show that the error term in Lemma 1(a) goes to 0 in probability uniformly over m.

Lemma 2.

(a)

$$\mathrm{E}I(D = m)\mathbf{X}^{\otimes l} \equiv \exp(\alpha_m) s_m^{(l)},$$

(b)

$$\frac{1}{N} \operatorname{Var} \sum_{i=1}^{N} I(D_i > 0) \mathbf{Z}_{D_i}^T \left\{ \mathbf{X}_i - \frac{s_{D_i}^{(1)}}{s_{D_i}^{(0)}} \right\} = \sum_{m=1}^{M} p_{1m} \mathbf{Z}_m^T \mathcal{J}_m \mathbf{Z}_m.$$

Proof of Lemma 2. The identity in part (a) easily follows as by standard conditional expectation arguments both sides of the identity can be shown to be equivalent to $\mathrm{E}\mathbf{X}^{\otimes l} \exp(\alpha_m + \mathbf{X}^T \boldsymbol{\beta}_m) \mathrm{P}(D = 0|\mathbf{X})$.

To prove part (b), first note that the variance term on the left side of the identity is given by

$$\frac{1}{N} \sum_{i=1}^{N} \mathrm{E}I(D_i > 0) \mathbf{Z}_{D_i}^T \left\{ \mathbf{X}_i - \frac{s_{D_i}^{(1)}}{s_{D_i}^{(0)}} \right\}^{\otimes 2} \mathbf{Z}_{D_i}$$

$$= \sum_{m=1}^{M} p_{1m} \mathbf{Z}_m^T \mathrm{E}\left[\left\{ \mathbf{X}_i - \frac{s_m^{(1)}}{s_m^{(0)}} \right\}^{\otimes 2} \middle| D = m \right] \mathbf{Z}_m.$$

Now, from the repeated use of identity (a),

$$\mathrm{E}\left[\left\{ \mathbf{X}_i - \frac{s_m^{(1)}}{s_m^{(0)}} \right\}^{\otimes 2} \middle| D = m \right]$$

$$= \frac{1}{p_{1m}} \mathrm{E}I(D = m) \left\{ \mathbf{X}_i - \frac{s_m^{(1)}}{s_m^{(0)}} \right\}^{\otimes 2}$$

$$= \frac{1}{\exp(\alpha_m) s_m^{(0)}} \left\{ \exp(\alpha_m) s_m^{(2)} - 2 \exp(\alpha_m) s_m^{(1)} \left[\frac{s_m^{(1)}}{s_m^{(0)}} \right]^T + \exp(\alpha_m) s_m^{(0)} \left[\frac{s_m^{(1)}}{s_m^{(0)}} \right]^{\otimes 2} \right\}$$

$$= \mathcal{J}_m$$

and, hence, part (b) is proved.

Proof of Proposition 1.

(a) *Consistency.* From the law of large numbers for triangular arrays,

$$\frac{1}{N} \mathbf{Z}^T T_{\boldsymbol{\beta}}^* \xrightarrow{P} \lim_{N \to \infty} \sum_{m=1}^{M} p_{1m} \mathbf{Z}_m^T \mathrm{E}\left\{ \mathbf{X}_i - \frac{s_m^{(1)}}{s_m^{(0)}} \middle| D = m \right\}. \tag{A.1}$$

Now, using the identity stated in Lemma 2(a), it is easy to show that each of the conditional expectations in expression (A.1) is 0. Thus, the main condition for consistency, the asymptotic unbiasedness of the score equations, is proved. Now, differentiation of the score functions gives

$$\frac{\partial}{\partial \boldsymbol{\theta}^T} \mathbf{Z}^T T_{\boldsymbol{\beta}}^* = \sum_{i=1}^{N} I(D_i > 0) \mathbf{Z}_{D_i}^T \left\{ \frac{S_{D_i}^{(2)}}{S_{D_i}^{(1)}} - \left[\frac{S_{D_i}^{(1)}}{S_{D_i}^{(0)}} \right]^{\otimes 2} \right\} \mathbf{Z}_{D_i}.$$

From the law of large numbers and condition (A.3),

$$\frac{\partial}{\partial \boldsymbol{\theta}^T} \{ \mathbf{Z}^T T_{\boldsymbol{\beta}}^* / N \} \xrightarrow{P} - \lim_{N \to \infty} \sum_{m} p_{1m} \mathbf{Z}_m^T \mathcal{J}_m \mathbf{Z}_m = -\mathcal{I}.$$

Using the boundedness conditions given in conditions (A.2) and (A.4), one can further show that the preceding convergence is uniform in $\boldsymbol{\theta}$ in an open neighborhood of $\boldsymbol{\theta}_0$. Consistency of $\{\hat{\boldsymbol{\theta}}_{\mathrm{PCL}}^N\}$ now follows from straight application of the results given in Foutz (1977).

(b) *Asymptotic normality.* To establish the given form of the asymptotic representation of the PCL estimator, first consider the Taylor series expansion

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathrm{PCL}}^N - \boldsymbol{\theta}_0\right) = -\left[\frac{1}{N} \frac{\partial}{\partial \boldsymbol{\theta}^T} \mathbf{Z}^T T_{\boldsymbol{\beta}}^* \right]^{-1} \frac{1}{\sqrt{N}} \mathbf{Z}^T T_{\boldsymbol{\beta}}^* + o_p(1).$$

Now write

$$\frac{1}{\sqrt{N}} \mathbf{Z}^T T_{\boldsymbol{\beta}}^* = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} I(D_i > 0) \mathbf{Z}_{D_i}^T \left\{ \mathbf{X}_i - \frac{s_{D_i}^{(1)}}{s_{D_i}^{(0)}} \right\}$$

$$- \frac{1}{\sqrt{N}} \sum_{i=1}^{N} I(D_i > 0) \mathbf{Z}_{D_i}^T \left\{ \frac{S_{D_i}^{(1)}}{S_{D_i}^{(0)}} - \frac{s_{D_i}^{(1)}}{s_{D_i}^{(0)}} \right\}.$$

Using Lemma 2, one can write the second term in the preceding expression as

$$\frac{1}{N} \sum_{i=1}^{N} I(D_i > 0) \mathbf{Z}_{D_i}^T \left[\frac{1}{\sqrt{N}} \sum_{j=1}^{N} I(D_j = 0) \times \frac{\exp(\mathbf{X}_j^T \boldsymbol{\beta}_{D_i})}{s_{D_i}^{(0)}} \left\{ \mathbf{X}_j - \frac{s_{D_i}^{(1)}}{s_{D_i}^{(0)}} \right\} \right] + o_p(1).$$

By changing the order of the two sums in the previous expression,

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{N} I(D_j = 0) \left[\frac{1}{N} \sum_{i=1}^{N} \frac{I(D_i > 0)}{s_{D_i}^{(0)}} \times \exp(\mathbf{X}_j^T \boldsymbol{\beta}_{D_i}) \mathbf{Z}_{D_i}^T \left\{ \mathbf{X}_j - \frac{s_{D_i}^{(1)}}{s_{D_i}^{(0)}} \right\} \right] + o_p(1).$$

By the law of large numbers, the expression within [ ] converges in probability to $\boldsymbol{\Gamma}_j$. This proves part (b) of the proposition. Asymptotic normality of the PCL estimator now follows from a standard application of the central limit theorem for triangular arrays. The given form of the asymptotic variance follows from part (b) of Lemma 2.

[Received May 2002. Revised June 2003.]

REFERENCES

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: Wiley.

Anderson, J. A. (1984), "Regression and Ordered Categorical Variables," *Journal of the Royal Statistical Society*, Ser. B, 46, 1–30.

Begg, C. B., and Zhang, Z. F. (1994), "Statistical Analysis of Molecular Epidemiologic Studies Employing Case-Series," *Cancer Epidemiology Biomarkers and Prevention*, 3, 173–175.

Dubin, N., and Pastermack, B. S. (1986), "Risk Assessment for Case–Control Subgroups by Polytomous Logistic Regression," *American Journal of Epidemiology*, 123, 1101–1117.

Foutz, R. V. (1977), "On the Unique Consistent Solution to the Likelihood Equations," *Journal of the American Statistical Association*, 72, 147–149.

Graubard, B. I., and Korn, E. L. (1987), "Choice of Column Scores for 2 × *k* Contingency Tables," *Biometrics*, 43, 471–476.

Greenland, S. (1994), "Alternative Models for Ordinal Logistic Regression," *Statistics in Medicine*, 13, 1665–1677.

Hosmer, D., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: Wiley.

Liang, K., Zeger, S., and Quaqish, B. (1992), "Multivariate Regression Analysis for Categorical Data," *Journal of the Royal Statistical Society*, Ser. B, 54, 3–40.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman & Hall.

Peters, U., Sinha, R., Chatterjee, N., Subar, A., Ziegler, R. G., Kulldorff, M., Bresalier, R., Weissfeld, J. L., Flood, A., Schatzkin, A., Hayes, R. B. for the PLCO Project Team (2003), "Dietary Fiber and Colorectal Adenoma in a Colorectal Cancer Early Detection Programme," *Lancet*, 361, 1491–1495.

Schroeder, J. C., and Weinberg, C. R. (2001), "Use of Missing Data Methods to Correct Bias and Improve Precision in Case–Control Studies in Which Cases Are Subtyped but Subtype Information Is Incomplete," *American Journal of Epidemiology*, 154, 954–962.

Terry, M. B., Gammon, M. D., Schoenberg, J. B., Brinton, L. A., Arber, N., and Hanina, H. (2002), "Oral Contraceptive Use and Cyclin D1 Overexpression in Breast Cancer Among Young Women," *Cancer Epidemiology Biomarkers and Prevention*, 11, 1100–1103.

van Der Vaart, A. W. (1996), *Asymptotic Statistics*, Cambridge, U.K.: Cambridge University Press.

Zhao, L., and Prentice, R. (1990), "Correlated Binary Regression Using a Quadratic Exponential Model," *Biometrika*, 77, 642–648.